DOCUMENT RESUME

ED 464 104 TM 033 785

AUTHOR King, Jason E.
TITLE Logistic Regression: Going beyond Point-and-Click.
PUB DATE 2002-04-00
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002). Contains small type.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algorithms; *Computer Software; Discriminant Analysis; Literature Reviews; *Regression (Statistics)
IDENTIFIERS *Statistical Analysis System; *Statistical Package for the Social Sciences

ABSTRACT

A review of the literature reveals that important statistical algorithms and indices pertaining to logistic regression are being underused. This paper describes logistic regression in comparison with discriminant analysis and linear regression, and suggests that some techniques only accessible through computer syntax should be consulted in evaluating logistic regression models. A heuristic dataset is used throughout the study to make the discussion complete, and logistic regression concepts are linked to the interpretation of results from a sample dataset. Statistical Analysis System and Statistical Package for the Social Sciences syntax files are provided for all the analyses described. The comparison shows the utility of logistic regression and the conditions under which it is more robust. Four appendixes contain the statistical package syntax examples. (Contains 4 tables, 4 figures, 6 endnotes, and 35 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

Running Head: LOGISTIC REGRESSION

ED 464 104

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. King

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Logistic Regression: Going Beyond Point-and-Click

Jason E. King

Baylor College of Medicine

Paper presented at the annual meeting of the American Educational Research Association, April, 2002.

Correspondence concerning this article should be addressed to Jason King, 1709 Dryden Suite 534, Medical Towers, Houston, TX. 77030. E-mail: Jasonk@bcm.tmc.edu

TM033785

BEST COPY AVAILABLE

Abstract

A search of the literature reveals that important statistical algorithms and indices pertaining to logistic regression are being underused. In layperson's terms, the writer (a) describes logistic regression in comparison with discriminant analysis and linear regression and (b) suggests that some techniques only accessible via computer syntax should also be consulted in evaluating logistic regression models. A heuristic dataset is employed throughout to make the discussion concrete. SAS and SPSS syntax files are provided for all analyses described.

Logistic Regression: Going Beyond Point-and-Click

Regression procedures are useful for understanding and explaining complex relationships among variables and for making predictions to a criterion. Linear regression models are regularly adopted for this purpose. While linear models are appropriate when both criterion (dependent) and predictor (independent) variables are continuously scaled, these should not be used when the criterion is categorically scaled. One reason for the avoidance is that predicted probabilities will often fall outside the permissible range. Further, the relationship between predictor and criterion will be underestimated if the underlying function is nonlinear.

An alternative to linear regression is discriminant analysis (DA). While DA is frequently the analysis of choice in the presence of a categorically-scaled criterion variable, it requires the strong assumptions of multivariate normality of the predictors and equality of covariance matrices (Klecka, 1980). When the assumptions are met, discriminant analysis may be a viable method (Rice, 1994), but this is not often the case. Further, DA cannot be (validly) applied with categorical predictor variables because multivariate normality cannot hold with non-continuous data.

In his book on regression methods Darlington writes, "[discriminant analysis] is in the process of being replaced in most modern practice by logistic regression" (1990, p. 458). Psychometricians regularly adopt logistic regression in estimating item response theory parameters (Hambleton & Swaminathan, 1985) and to assess differential item functioning (French & Miller, 1996; Swaminathan & Rogers, 1990). Logistic regression has proven to be especially useful in epidemiology (Lemeshow & Hosmer, 1982), family studies (DeMaris, 1995; Morgan & Teachman, 1988), and in other more specialized areas where criterion variables are often dichotomous (e.g., see Schiel & King [1999] for a recent example of its application in making course placement decisions). In addition, researchers from various fields have recently published "tutorials" encouraging greater use of logistic regression (e.g., clinical psychology: Davis & Offord, 1997; counseling psychology: Cizek & Fitzgerald, 1999; health care: Peng, Manz & Keck, 2001; interpersonal violence: McNutt, Holcomb & Carlson, 2000; social work: Morrow-Howell & Proctor, 1992; and sociology: Lottes, Adler & DeMaris, 1996).

There are several reasons for the increased popularity of the statistical technique in these disciplines. Most notably, logistic regression is a more general technique than other methodological choices. The related procedure of loglinear analysis can accept a qualitative criterion but not continuous predictors. Discriminant analysis allows a qualitative criterion but not categorical predictors. Logistic regression, on the other hand, allows a qualitative criterion and predictors that are continuous, categorical or a mixture of both.

Second, logistic regression fits a nonlinear function to the data. The relationship between predictors and a dichotomous criterion[1] is nonlinear and will not be adequately modeled by linear regression. When the canonical discriminant functions are used for prediction, discriminant analysis also supposes a linear relationship between criterion and predictors. Logistic regression models a curvilinear function to the data, which potentially will explain more criterion variance.

Third, the restrictive assumptions of linear regression and discriminant analysis are relaxed in logistic regression. In addition, linear regression will not necessarily yield predicted scores between 0 and 1, but logistic regression always computes permissible probabilities.

Purpose

This paper explains the statistical method of logistic regression by drawing analogies, when possible, to concepts and calculations associated with the more familiar linear regression and discriminant analysis methods. Logistic regression concepts are also linked directly to the interpretation of results from a sample dataset. Computer application is made to both SAS and

SPSS. When important procedures are not readily available in the commercial packages, syntax files are appended to facilitate greater use of these techniques.

Why Not Use Linear Regression?

Calculating the Estimates

A brief review of the workings of linear regression will aid in understanding the dynamics underlying discriminant analysis and logistic regression. In linear regression one derives an equation composed of predictor variables that maximally explains the variation of scores on the dependent variable. When several predictors are included, the equation is composed of multiplicative constants (b weights) applied to scores on the independent variables, along with a single additive constant (a weight). When an error term (e) is included, the equation perfectly defines each individual's criterion score:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e \quad , \qquad (1)$$

where k = the number of predictor variables. If the error term is excluded, scores resulting from the equation produce a synthetic variable composed of predicted criterion scores, denoted $\hat{Y}$:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k \quad . \qquad (2)$$

More will be said about predicted scores later.

Unstandardized regression weights are derived such that the sum of the squared deviations of the criterion scores (Y) from the predicted ($\hat{Y}$) scores are minimized. This is equivalent to saying that the (squared) errors are minimized, hence the mathematical method of obtaining such a solution is denoted as ordinary least squares (OLS) estimation. No other solution will produce smaller (squared) errors if the assumptions of linear regression are met.

The a weight (additive constant, intercept) is interpreted as the criterion score when the values of the predictor variables are at 0. The b weight (multiplicative constant, slope) indicates the change in the mean of the probability distribution of Y per unit increase in X. In other words, if the predictor variable is increased by 1 point and all other predictors are held constant, how many raw score units will the criterion variable increase?

When all scores are standardized (i.e., transformed to have variance of 1 and mean of 0), the a weight disappears (becomes 0) and the b weights are denoted as β weights or standardized regression coefficients. They are interpreted: If a standardized predictor variable is increased by 1 unit while holding all other predictors constant, how many standardized units will the criterion variable increase? Larger β weights indicate stronger predictors in the equation. However, because multicollinearity may exist among predictors, one should also always interpret structure coefficients (or bivariate correlations) when determining variable importance (Thompson & Borrello, 1985).

Assumptions

Linear regression assumes that (a) each X variable is measured without error, (b) the relation between Y and each X is linear (in the parameters), (c) the mean of errors is 0, (d) errors are uncorrelated (independence of observations), (e) error variance is constant across levels of each X (homoscedasticity), (f) errors are uncorrelated with each X, (g) errors are normally distributed, (h) Y is a random variable, and (i) no important

variables are excluded from the equation, thus implying a correctly specified model. Assumption (b) can be relaxed by transforming (raising to powers, etc.) the predictors or criterion variables, but the parameters themselves (i.e., the a and b weights) cannot be so transformed. In this sense, the equation is linear with respect to the parameters, but not necessarily with respect to the variables. This distinction will be important when considering logistic regression. Assumption (h) suggests that Y has a range of possible values, each having an associated probability. If the criterion is categorical, then assumptions (b), (e), (g), and (h) will also necessarily fail. Thus, linear regression should not be used with a categorical criterion variable or categorical predictor variables (unless dummy or effect coding is applied).

To see these dynamics, consider the sample data presented in Table 1. These are taken from the cars dataset included in several recent versions of the SPSS statistical package. You may wish to replicate the following analyses if you have access to SPSS. Cases with any missing values were first deleted, next, the year and cylinder categorical variables were deleted, and finally, country of origin was recoded to a dichotomy: 0 = European/Japanese country of origin, 1 = American origin. This resulted in a total of N = 391 usable cases measured on six interval/ratio-scaled variables.

Using linear regression, the country of origin in which each car was manufactured was regressed on miles per gallon (mpg), engine size, horsepower, vehicle weight, and time to accelerate. Figure 1 depicts a scatterplot of one of the predictors (mpg) with the criterion. Note that the scores are not clustered around a diagonal line, as one would see if the relationship between the variables was linear. Assumption (b) will not be met because of the dichotomous dependent variable.

A Multiple R^2 of .495 was obtained from the linear regression (see Table 2). Statistical significance tests computed for each predictor's b weight indicate that mpg, engine, and horsepower are the most important predictors in the equation, though structure coefficients point to weight as a good predictor of origin as well.[2] Figures 2 and 3 demonstrate how the qualitative dependent variable caused failure to meet the error distributional assumptions. The errors are not normally distributed (Figure 2; plotted points are expected to randomly cluster around the diagonal line) nor identical in variance across values of the predictor variable (Figure 3 depicts mpg; a random pattern of plotted points is expected). Clearly, linear regression should not be used with this model.

Even if one argues that the failure to meet those assumptions does not invalidate results, consider the predicted scores obtained using Equation 2. For this analysis, the $\hat{Y}$ scores ranged from .06 to 1.42 (not presented here). With a dichotomous criterion variable, the $\hat{Y}$ scores are equivalent to predicted probabilities of group membership equal to 1, in this case, a car being manufactured in America. However, a probability cannot fall outside the 0 to 1 range. What does it mean to say that for certain combinations of scores on predictor variables, the probability of being manufactured in America is 1.42? One might argue that such a car is predicted with certainty to be in group 1. However, predicted scores closer to the typical cutting point of .5 can be invalid as well, in which case group prediction will be incorrect. This illustrates a second problem with using linear regression, namely, for some vectors (combinations) of independent variable scores, predicted criterion scores (probabilities) will be invalid.

Why Not Use Discriminant Analysis?

Calculating the Estimates

Discriminant analysis (DA) holds more promise when analyzing categorical variables (for a full treatment, see Klecka, 1980; Stevens, 1996). DA can also be used for either description or prediction. This analysis requires a qualitative criterion (grouping variable) along with continuously-scaled predictors. In DA one or more canonical discriminant functions are created by

linearly combining the discriminating (predictor) variables to maximally discriminate between values on the dependent variable (i.e., groups). Although the criterion variable is assumed to be categorical in this analytic approach, the technique is similar to linear regression. Both linear regression and DA use OLS estimation in deriving the equation/function (equations for DA will not be presented). Consequently, additive and multiplicative weights are created which are analogous to *a* and *b* weights in regression. In this context, *b* weights are referred to as unstandardized canonical discriminant function coefficients. As in linear regression, both structure coefficients and predicted scores can also be obtained.

The two statistical procedures differ in that DA allows for estimation of multiple functions (equations) to better separate scores on the criterion variable. The number of canonical discriminant functions that can be estimated is equal to *k* - 1, where *k* = the number of predictor (discriminating) variables.[3] Each function can then be evaluated for statistical and substantive significance. For our sample data containing only two groups, namely European/Japanese versus American country of origin, only one function (equation) can be estimated.

Table 3 presents DA results using the cars data to estimate the model described earlier. Again, mpg, engine and horsepower are the strongest discriminators in the equation as evidenced by their relatively large standardized canonical discriminant function coefficients. Because DA is a multivariate procedure, a multivariate measure of effect size, Wilks' λ, is applied. Subtracting λ from 1 yields a Multiple R^2 measure of effect size, here 1 - .505 = .495. Note that this value is identical to that obtained through linear regression indicating that DA does not model the non-linear relationship any better.[4] Recall also that the linear regression analysis yielded unacceptable probabilities for the sample data. Predicted probabilities realized through discriminant analysis ranged from .01 to .99; all within the permissible range, as will always be the case with DA.

Assumptions

Discriminant analysis requires strict assumptions. Klecka (1980) lists three: (a) no variable may be a linear combination of other variables, (b) each group must be drawn from a population that is multivariate normal, and (c) population covariance matrices must be equal for each group. Stevens (1996) adds and emphasizes the importance of meeting the assumption of (d) independence of observations. Though not always explicitly stated in statistics books, as in linear regression DA assumes that (e) each *X* is measured without error, (f) the mean of errors is 0, (g) errors are uncorrelated with each *X*, and (h) no important variables are excluded from the equation.

Regarding assumption (a), engine and weight are correlated at .934. This value indicates likely multicollinearity. Assumption (b) should be assessed by first evaluating bivariate normality (e.g., via scatterplots) and then multivariate normality (see Thompson, 1990, for suggested methods). But with a qualitative predictor, the assumption cannot hold because a qualitative variable is not normally distributed and will not yield bivariate normal distributions. For the cars data with no categorical predictors, the assumption may be tenable (based on analyses not presented here). Box's test of equality of covariance matrices can be used to evaluate assumption (c). For these data the assumption fails, $F(15,382531) = 32.524$, $p < .001$. Therefore, DA will likely produce invalid results.

Logistic Regression

Calculating the Estimates

Unaltered criterion, transformed predictors. Because of the nonlinear function obtained when predicting to a dichotomous dependent variable, the logistic function (and its associated equation) differs dramatically from

linear regression. Figure 4 depicts a typical function derived in a single predictor model.

The multiple predictor linear regression equation (Equation 1) differs significantly from the equation used in logistic regression:

$$\underline{Y} = \frac{\exp(\underline{a} + \underline{b}_1\underline{X}_1 + \underline{b}_2\underline{X}_2 + \ldots + \underline{b}_k\underline{X}_k)}{1 + \exp(\underline{a} + \underline{b}_1\underline{X}_1 + \underline{b}_2\underline{X}_2 + \ldots + \underline{b}_k\underline{X}_k)} + \underline{e} \quad . \qquad (3)$$

Instead of linearly relating the predictor scores to the criterion, an exponential function of the predictors is modeled. The equation is no longer "linear with respect to the parameters," as was the case in linear regression. Because of the curvilinear function, the slope is interpreted differently as well. In linear regression the amount $\underline{Y}$ is increased remains constant across the function, while in logistic regression for a unit increase in $\underline{X}$, the amount that $\underline{Y}$ increases will vary depending on where the $\underline{X}$ value falls along the function. Consider again Figure 4: $\underline{X}$ values in the tails are associated with smaller increases in $\underline{Y}$.

Regarding predicted scores, $\hat{\pi}$ instead of $\hat{\underline{Y}}$ represents the logistic regression expected conditional mean (predicted score, predicted probability) of the dependent variable given a certain combination of scores on the predictor variables. The equation is defined as

$$\hat{\underline{Y}} = \hat{\pi} = \frac{\exp(\underline{a} + \underline{b}_1\underline{X}_1 + \underline{b}_2\underline{X}_2 + \ldots + \underline{b}_k\underline{X}_k)}{1 + \exp(\underline{a} + \underline{b}_1\underline{X}_1 + \underline{b}_2\underline{X}_2 + \ldots + \underline{b}_k\underline{X}_k)} \quad . \qquad (4)$$

Recall that with a dichotomous criterion variable, $\hat{\underline{Y}}$ scores in linear regression and DA are simply the probability of group membership being equal to 1. $\hat{\pi}$ scores in logistic regression are interpreted identically.

Logistic regression is often performed using a maximum likelihood algorithm to estimate the parameters rather than OLS (for details, see Neter, Kutner, Nachtsheim, & Wasserman, 1996, p. 573ff), though other options are available. This procedure is iterative so that various values for the $\underline{a}$ and $\underline{b}$ parameters are tested until a best-fitting solution is found (i.e., one that maximizes the log-likelihood function). Pedhazur explains, "In logistic regression the aim is to estimate parameters most likely to have given rise to the sample data. Hence, the name maximum likelihood..." (1997, p. 718). A disadvantage of such iterative procedures, and hence of logistic regression in general, is the possibility that the solution will not converge on a best estimate.

Transformed criterion, unaltered predictors. In Equations 3 and 4 the predictor variables are raised to an exponent. The logistic function can alternatively be defined such that instead of exponentiating the predictors while leaving the criterion unaltered, the criterion is transformed while leaving the predictors unaltered. But the criterion must first be expressed in terms of odds, not probabilities. The odds of an event occurring is[5] defined as the probability that the event will occur (π) divided by the probability that it will not occur ($1 - \pi$). The observed probability of a criterion response of 1 for a given vector (combination/set) of independent variable scores can be determined by aggregating all the cases having that particular vector of predictor scores and then simply calculating the percentage of criterion responses equal to 1. The odds is then obtained by dividing the percentage of responses equal to 1 (i.e., π) by the percentage of responses not equal to 1 (i.e., $1 - \pi$).

Odds and probability are easily confused, so an example is needed.[6] Using our sample data, suppose we predict origin using only mpg. To determine the observed probability of a criterion response of 1 occurring for cars with an mpg score of, say 24, simply count the number of origin scores equal to 1 for every car having that mpg value: 6 of 18 for these data. Thus, based on

our data, the observed probability that a car averaging 24 miles per gallon has been made in America is 6 / 18 = .33. The probability is 1 - .33 = .67 that the car was not made in America (and thus made in Europe or Japan).

Next, the odds is .33 / (1 - .33) = .5 that such a car was produced in the U.S. Odds less than 1.0 indicate that the event is less likely to occur than the absence of the event, and vice-versa. The odds of .5 in this case is interpreted: for any car having an mpg of 24, the car's origin is half as likely to be American as European/Japanese. If the odds was found to be 2.0, the interpretation would be: cars with an mpg score of 24 are twice as likely to have been made in America.

After transforming the data into odds, the natural logarithm must be taken. Just as subtraction negates addition, the natural logarithm ("ln" or "$\log_e$" on a calculator) negates an exponent ("e," "exp," or "inv+ln" on a calculator). So, a natural logarithm transformation is applied to the criterion scores in Equation 4 to negate the exponential predictor variables. This results in a "log-odds" or a "logit" for each case in the sample.

The right-hand side of Equation 4 can now be expressed in linear terms:

$$\ln \frac{\hat{\pi}}{1 - \hat{\pi}} = \underline{a} + \underline{b}_1\underline{X}_1 + \underline{b}_2\underline{X}_2 + \ldots + \underline{b}_k\underline{X}_k \quad . \qquad (5)$$

This equation is linear with respect to the logits and is denoted the fitted logit response function (Neter et al., 1996). Logistic regression is often referred to as "logit" modeling due to the transformation of the dependent variable into linear log-odds/logits. These equivalent formulas (Equations 4 and 5) are presented merely to facilitate ease of interpretation. Menard (1995) emphasizes,

> It is important to understand that the probability...and the logit are [two] different ways of expressing exactly the same thing [the curvilinear relationship between criterion and predictors]. Of the [two] measures, the probability [transformed predictor variables]...is probably the most easily understood. Mathematically, however, the logit form of the probability [transformed criterion] is the one that best helps us to analyze dichotomous dependent variables [primarily in terms of interpreting variable importance]. (p. 13)

Assumptions

Unlike linear regression, logistic regression does not assume homoscedasticity or normality of errors. Unlike discriminant analysis, the assumptions of equality of covariance matrices and multivariate normality are not required. However, meeting these two assumptions will usually produce more stable parameter estimates. As in linear regression and DA, logistic regression assumes that (a) each $\underline{X}$ is measured without error, (b) the mean of errors is 0, (c) errors are uncorrelated [independence of observations], (d) errors are uncorrelated with each $\underline{X}$, and (e) no important variables are excluded from the equation. Also, the relationship between logits and predictors should be linear and additive (Fuller, 1998), though this requirement can be relaxed through transformation (see Neter et al., 1996, for a fuller discussion).

Computer Software Used to Obtain Logistic Estimates

Most of the popular statistical software packages now include logistic regression algorithms. Prior to the release of Version 6.06, SAS users were required to use PROC CATMOD, a general procedure for modeling categorical data, to obtain logistic results. The newer versions include a procedure dedicated to logistic regression. Appendix A presents examples of SAS and SPSS syntax statements for conducting a logistic regression, although Windows versions of both applications allow users to "point-and-click" to obtain some analyses.

Keep in mind that statistical packages may differ as to the default

criterion score predicted. For the cars data, SAS (Version 8.0) predicted a criterion response of 0 (i.e., the absence of the event), while SPSS (Version 10.1) predicted a response of 1. This will occasion some estimates to diverge, most notably odds ratios (see discussion below), unless criterion scores are re-coded before running the analysis.

Table 4 depicts results from a logistic regression. SAS was used to estimate the parameters. SPSS estimates would be identical but with reversed signs due to predicting origin = 1 instead of 0.

Assessing Overall Model Fit

There are several ways to assess overall model fit in logistic regression. While various statistical significance tests may be consulted, it is preferable to always interpret significance tests in light of effect size measures (on the significance testing controversy, see Carver, 1993; Thompson, 1999; Wilkinson & The APA Task Force on Statistical Inference, 1999). Several tests of statistical significance and measures of effect size will be described and interpreted in the context of our sample data.

Statistical significance tests. Significance tests in logistic regression predominantly employ chi-square distributions to obtain probability values. One common index, denoted "-2LL," is computed by first multiplying the log likelihood (LL) obtained from a model containing all predictors by -2, thus producing a statistic which is chi-square distributed. A second chi-square distributed statistic is computed for an intercept-only null model. The difference between the chi-square values is itself chi-square distributed and can be tested for statistical significance. This procedure is analogous to an overall F-test in linear regression, which evaluates the hypothesis that all predictors are related to the dependent variable (i.e., $b_1 = b_2 = \ldots = b_k = 0$). A small p value indicates that at least one predictor is related to the criterion, hence statistical significance is desired in this case.

In SPSS the null and full model -2LL values are not listed unless the iteration history is requested. However, by default the chi-square difference between the two log likelihoods is listed under the title "Omnibus Tests of Model Coefficients" and on the "Model" row.

In SAS both -2LL values are depicted under the title "Model Fit Statistics" and on the "-2 LOG L" row. The null model -2LL is labeled "Intercept Only," while the full model -2LL is labeled "Intercept and Covariates." The chi-square difference is then provided in the section entitled "Testing Global Null Hypothesis: Beta = 0" and on the "Likelihood Ratio" row.

For the sample data the following values were obtained: Null Model -2LL = 517.724, Full Model -2LL = 212.659, Difference $\chi^2(5) = 305.065$, $p < .001$ (see Table 4). The statistically significant chi-square difference implies that the five independent variables are at least somewhat predictive of country of origin. Other related significance tests are available as well (e.g., the Score statistic, the Akaike Information Criterion (AIC), the Schwartz criterion).

One may also test the statistical significance of each predictor in any model of interest. This is accomplished using either the Wald test or a leave-one-variable-out technique. The latter procedure entails comparing the -2LL obtained for a model that includes the predictor variable of interest to the -2LL for a model excluding the variable. The difference in log likelihoods is distributed as a chi-square variate with 1 degree of freedom. While this procedure is slightly more accurate than the Wald test, the two will generally produce similar results. Both SPSS and SAS print Wald statistics for each variable modeled (see Table 4).

Effect size measures. Effect size indices quantify the strength of association between variables after removing the effect of sample size. SPSS print two R^2-type measures of effect size. The Cox and Snell R^2 (here, .504)

is calculated as R^2 = 1 - [Null Model -2LL (-) Full Model -2LL]$^{2/N}$, where N = sample size (SPSS, Inc., 1999). Because this R^2 measure cannot achieve 1.0 by definition, another option is to calculate the Nagelkerke R^2 (equation not presented here). SAS provides both indices as well, but labels them as generalized R^2 and max-rescaled R^2, respectively. The two statistics are not available through the "analyst" application, which is a point-and-click environment, but must be requested via SAS code (SAS Institute, Inc., 1999). Specifically, one must add /RSQUARE to the end of the MODEL statement in PROC LOGISTIC.

The most popular R^2-type measure in logistic regression is obtained by dividing the difference between the null and full model -2LL values by the null model -2LL value: (Null Model -2LL (-) Full Model -2LL) / Null Model -2LL. This index is denoted R_L^2 by Hosmer and Lemeshow (1989). Neither SAS nor SPSS include this useful index of the proportional reduction in the null model log likelihood attributable to the predictors, but it can easily be computed by hand. For the cars data we obtain: R_L^2 = (517.724 - 212.659) / 517.724 =.589 A version of R_L^2 "adjusted" for the number of predictors in the model can be found as follows: $R_{L\text{-}adj}^2$ = ((1 - Null Model -2LL (-) Full Model -2LL) - 2k) / Null Model -2LL, where k = number of predictors in the full model. Here, it would be $R_{L\text{-}adj}^2$ = .489.

While these indices do assess effect size, they are not variance-accounted-for indices because the log likelihood is not really a sum of squares. However, it is possible to calculate a measure in logistic regression which is analogous to the R^2 obtained in linear regression (see, e.g., Menard, 1995). Again, though the calculations are straight-forward, the popular statistics packages do not include this index. First, calculate the logistic regression estimates saving the predicted values. Next, use a linear regression routine to predict the criterion scores from the predicted scores obtained from the logistic regression (see Appendix B for SPSS syntax, though the same can be generated in SAS). The obtained R^2 indexes the proportion of variance in the dependent variable accounted for by the logistic equation (equivalently, one could have correlated the criterion and predicted scores to obtain the Multiple R). For the cars data, an R^2 of .622 was obtained. Note the larger percentage of variance explained when using logistic regression (62.2%) as opposed to linear regression (49.5%) and discriminant analysis (49.5%).

Other techniques. Other procedures for evaluating model fit include those designed to compare the proportion of cases correctly predicted by the model (i.e., predicted to be in group 0 or 1) with observed criterion scores. For example, the c statistic is printed in SAS, along with several other measures of concordance. Classification tables and histograms of estimated probabilities can also be useful for determining optimal cutoff points. Histograms may be inspected to determine whether a rule other than the default (.5) for assigning cases to groups might be more valid for a given purpose.

As an example of the use of contingency tables, consider a statistic proposed by Lemeshow and Hosmer (1982). A chi-square distributed statistic is calculated by summing squared residuals. Each residual is obtained by taking the difference between the observed and expected criterion score and dividing by the expected criterion score. This is done for all data points falling in each of a number of categories that were created by ranking the predicted probabilities (for details, see Lemeshow and Hosmer). In SPSS this index is provided by clicking on the "Hosmer & Lemeshow test" option, though SAS does not yet allow for its calculation. A small chi-square value is ideal (hence statistical significance is not desired in this case), signifying small differences between predicted and obtained scores. Our sample data produced the following results: $\chi^2(8)$ = 3.121, p = .927. This non-statistically significant finding suggests that the criterion variable is adequately explained by the five-predictor model.

Selecting Optimal Predictors

Researchers generally adopt regression models either to evaluate the

tenability of theories or to make predictions. Infrequently, only a single model may be estimated and evaluated. More often, researchers test and modify several competing theories based on the results obtained. At times, one is not concerned with the selection of predictor variables from a theoretical stance, but only with how well the equation predicts the criterion (e.g., distinguishes between two or more groups). In either case, analysts wish to pare down their original group of variables to a smaller group of variables which meet their theoretical or prediction standards.

Several options are available for locating optimal predictors. One may can inspect the weights and structure coefficients given to each variable in the regression equation [see below]. Second, one may calculate the relative percentage of variance explained by various combinations of variables entered as blocks in a series of analyses. Third, procedures can be applied which ostensibly select the "best" predictors (i.e., stepwise-type methods). But the problems associated with stepwise-type variable selection procedures have been documented elsewhere (see Huberty, 1989; Pedhazur, 1997, p. 211ff; Thompson, 1989). In spite of these warnings, both SAS and SPSS only include stepwise-type selection procedures (i.e., forward, backward, stepwise).

A better option is to calculate an "all possible" or "best subsets" logistic regression procedure. This method presents a summary statistic (e.g., R^2 value) for every possible combination of predictor variables, thus allowing the researcher to decide on substantive grounds as to the optimal predictor set. Although Hosmer, Jovanovic and Lemeshow (1989) and Hosmer and Lemshow (1989) have described how a best-subsets model selection procedure could be accomplished within logistic regression, no examples using SAS or SPSS computer syntax have been provided in any source to our knowledge (but see King, under review, for a full explanation).

Hosmer et al. (1989) recommend interpreting Mallow's measure of predictive squared error (i.e., Mallow's C_p) for identification of a best subset of variables. C_p is expected to be approximately equal to $1 + p$, where p = the number of predictors in the reduced model, with smaller values preferred.[7]

Appendixes C and D provide the necessary SAS syntax and results using the cars data. Note that the procedure is run under a linear regression routine but with a transformed dependent variable (z) and a case weight (u). In this case, the C_p of smallest magnitude relative to p was obtained for the three-variable model consisting of engine, horsepower, and weight, $C_p = 2.498$. The expected C_p was 3 (variables) + 1 = 4. Thus, if one wished to select the best subset of predictor variables according to model parsimony, these would be chosen. Of course theory should also drive model selection unless prediction is of sole interest.

Interpreting Model Parameters

Transformed criterion, unaltered predictors. After a model has been selected, one moves to interpretation of the estimated parameters. From the coefficients listed in Table 4, the estimated prediction equation for the cars data (in logits) is

$$\ln \frac{\hat{\pi}(\text{orig} = 1)}{1 - \hat{\pi}(\text{orig} = 1)} = -.114 + .035\text{mpg} - .105\text{eng} + .050\text{horse} + .004\text{weight} + -.028\text{accel} \quad . \tag{6}$$

As explained earlier, in logistic regression the unstandardized b weight represents the change in criterion variable logits (i.e., the logarithm of the odds associated with the criterion variable) for a 1-unit increase on the predictor variable. For example, from Table 4 horsepower was given a b of -.050. This is interpreted as: a 1-unit increase in horsepower produces a .050 increase in origin log-odds (logits). Standardized β weights are also printed in SAS (but not SPSS) and are useful for comparing strength of prediction across variables having dissimilar standard deviations.

To understand how to interpret b weights, assume one wishes to predict the country of origin for a car with the following dimensions: 25 mpg, 100 cu in. engine displacement, 90 horsepower, 2000 lbs weight, and 20 s to accelerate from 0 to 60 mph. Using Equation 4, the a and b coefficients from Table 4, and an Excel spreadsheet, one can calculate the predicted probability of origin being equal to 0 (European/Japanese-made): $\hat{\pi}$ = "=EXP(-.1136+(.0349*25)+(-.1054*100)+(.0504*90)+(.0035*2000)+(-.0280*20))/(1+EXP(-.1136+(.0349*25)+(-.1054*100)+(.0504*90)+(.0035*2000)+(-.0280*20)))" = .768. Given this prediction model, it is likely ($\hat{\pi}$ = .768) that a car with those particular dimensions would be foreign made; it is less likely ($\hat{\pi}$ = 1 - .768 = .232) that the car would be made in America. The predicted odds of the car being foreign is $\hat{\pi}$ / (1 - $\hat{\pi}$) = .768 / (.232) = 3.303. In words, a car with these dimensions is 3 times more likely to be foregin. Taking the natural logarithm, one obtains a predicted log-odds (logit) of 1.195.

Now consider a car with exactly the same dimensions but having a 1-unit higher level of horsepower (i.e., 91). The predicted probability of the car being foreign (using the Excel function) is now .776, with a predicted odds of 3.474, and predicted log-odds of 1.2453. The difference between these two logits is 3.474 - 3.303 = .050. Thus, a 1-unit increase in horsepower resulted in a .050 increase in country of origin logits, which exactly matches the value of the b coefficient for horsepower listed in Table 4. Thus, holding all other variables in the question constant, a car having higher horsepower is more likely to be foreign made.

Transformed criterion, transformed predictors. Interpretation of log-odds change is difficult. What does it mean to say that increasing horsepower increases the logits of country of origin by .050? Equation 6 can be exponentiated to remove the logarithm, converting the dependent variable from logits back to odds, yet not all the way back to a probability. This model falls in between the two presented earlier (see Equations 4 and 5). The equation becomes

$$\frac{\hat{\pi}(orig = 1)}{1 - \hat{\pi}(orig = 1)} = e^{-.163} e^{-.021mpg} e^{.102eng} e^{-.047horse} e^{-.004weight} e^{.018accel} . \qquad (7)$$

Interpretation of the weights in this form is now somewhat clearer. The value to the left of the equals sign is just the odds of country of origin being equal to 0. A change in 1 unit for each predictor variable multiplies the odds of the criterion by e^b. So now we focus not on how many logits the criterion will increase additively (as we did earlier in subtracting two odds to obtain .050), but on how the odds of the criterion will increase multiplicatively (actually we will divide two odds, and division is the inverse of multiplication).

An example is needed, but first some terminology. The ratio of two odds for a 1-unit change in the predictor is termed, not surprisingly, an odds ratio. SAS labels it so, but SPSS denotes the odds ratio as Exp(B). If the ratio is greater than 1, the odds or likelihood of the event happening is increased; and the converse is true. If the odds ratio is 1.5, the odds of the event happening has increased 50%. An odds ratio of 1.0 indicates that the odds of the event has not changed, and the predictor is not related to the criterion.

Earlier we obtained the predicted odds for horsepower values of 90 (odds = 3.303) and 91 (odds = 3.474). The ratio of the two odds is 3.474 / 3.303 = 1.052. This is the multiplicative value by which the odds of the criterion will increase if horsepower is raised by 1 unit (note that this exactly matches the OR for horsepower in Table 4). It is interpreted: for every 1-unit increase in horsepower and holding all other variables constant, the odds of a car being foreign-made is increased by about 5%. This interpretation is more straight-forward than that based on logits, hence the popularity of odds

ratios in logistic regression. SPSS includes confidence intervals for each predictor's odds ratio. If a given interval spans 1, the hypothesis of no relation between predictor and criterion cannot be rejected.

For another example, what if horsepower were decreased from 90 to 75? Using Equation 4 and a spreadsheet, the odds ratio is now .446. Decreasing horsepower by 15 points makes a car (1 - .446 =) 55.4% less likely to be foreign-made; equivalently, the odds has decreased 55.4%. This does not imply that the odds of the car is now more likely to be American-made, only that the odds, whatever the value, has decreased 55.4% relative to its previous value. In fact, the odds is still relatively high (1.475) that a car having those dimension is foreign-made, but the odds has decreased from 3.303 when horsepower was set at 90. So the odds has been approximately halved by dropping horsepower 15 points.

Conclusion

A comparison among linear regression, discriminant analysis and logistic regression suggests the utility of the latter when appropriate. While DA often yields greater asymptotic relative efficiency (Bull & Donner, 1987; Efron, 1975), logistic regression is more robust due to the restrictive assumptions of the OLS estimation method (Press & Wilson, 1978). Some have cautioned:

> It is unlikely that the two methods will give markedly different results, or yield substantially different linear functions unless there is a large proportion of observations whose x-values lie in regions of the factor space with linear logistic response probabilities near zero or one. (Press & Wilson, p. 705)

Nevertheless, an additional 12% variance was explained using logistic regression when applied to the arbitrarily-selected sample dataset. Further, due to the violation of the multivariate normality assumption inherent when categorical predictors are present, logistic regression should be the analysis of choice for most research studies involving a qualitative criterion. With a model that posits more realistic nonlinear relationships among variables, educational researchers can achieve enhanced statistical precision in estimating prediction equations, which should facilitate improved decision making.

Footnotes

[1]Polytomous criterion variables will not be considered in this paper.

[2]Multicollinearity has produced this effect. Vehicle weight is highly correlated with other good predictors: engine displacement (r_{xx} = .934), horsepower (r_{xx} = .863) and mpg (r_{xx} = -.831). In fact, although horsepower has a smaller structure coefficient and a smaller bivariate correlation with the criterion (r_{xy} = -.489) than does vehicle weight (r_{xy} = .601), horsepower is less correlated with engine displacement (r_{xx} = -.898) and mpg (r_{xx} = -.776), thereby explaining more unique criterion variance. Consequently, horsepower is given greater weight in the equation.

[3]On a side note, some prefer to use Fisher's classification functions to classify cases in predictive discriminant analysis. In that event, k classification functions are created, one for each group. Cases are classified into the group on which they obtain the highest score. We focus here on the discriminant functions because, "while some cases may be classified differently in this instance [when using classification functions], the canonical discriminant function results should be more accurate, because the effect of idiosyncratic sample variation has been reduced [through selecting a subset of functions]" (Klecka, 1980, p. 48).

[4]Classification plots are also useful in evaluating model fit, but will not be discussed here.

[5]In this context, "odds" takes singular verbs: "the odds is 2.0," not "the odds are 2.0."

[6]Here we are simply illustrating the dynamics of odds and probabilities, in this case observed probabilities. We are not demonstrating the actual computations involved in obtaining logistic regression parameter estimates.

[7]Hosmer et al. (1989) equivalently denote C_p as C_q.

References

Bull, S. B., & Donner, A. (1987). The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. Journal of the American Statistical Association, 82, 1118-1122.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cizek, G. J., & Fitzgerald, S. M. (1999). An introduction to logistic regression. Measurement and Evaluation in Counseling and Development, 31(4), 223-244.

Darlington, R. B. (1990) Regression and linear models. New York: McGraw-Hill.

Davis, L. J., & Offord, K. P. (1997). Logistic regression. Journal of Personality Assessment, 68, 497-507.

DeMaris, A. (1995). A tutorial in logistic regression. Journal of Marriage & the Family, 57, 956-968.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association, 70, 892-899.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. Journal of Educational Measurement, 33, 315-332.

Fuller, D. K. (1998, April). Sample size inequality and assumption violation in logistic regression. Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans, LA.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Hosmer, D. W., Jovanovic, B., & Lemeshow, S. (1989) Best subsets logistic regression. Biometrics, 45, 1265-1270.

Hosmer, D. W., & Lemeshow, S. (1989). Applied logistic regression. New York: John Wiley & Sons.

Huberty, C. J. (1989). Problems with stepwise methods--better alternatives. Advances in Social Science Methodology, 1, 43-70.

King, J. E. (under review). Running a best subsets logistic regression: An alternative to stepwise methods.

Klecka, W. R. (1980). Discriminant analysis. (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-119). Newbury Park, CA: Sage.

Lemeshow, S. & Hosmer, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology, 115, 92-106.

Lottes, I. L., Adler, M. A., & DeMaris, A. (1996). Using and interpreting logistic regression: A guide for teachers and students. Teaching Sociology, 24, 284-298.

Mallows, C. L. (1973). Some comments on C_p. Technometrics, 15, 661-675.

McNutt, L. A., Holcomb, J. P., & Carlson, B. E. (2000). Logistic regression analysis: When the odds ratio does not work: An example using intimate partner data. Journal of Interpersonal Violence, 15, 1050-1059.

Menard, S. (1995). Applied logistic regression analysis (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-106). Thousand Oaks, CA: Sage.

Morgan, S. P., & Teachman, J. D. (1988). Logistic regression: Description, examples, and comparisons. Journal of Marriage and the Family, 50, 929-936.

Morrow-Howell, N, & Proctor, E. K. (1992). The use of logistic regression in social work research. Journal of Social Service Research, 16(1-2), 87-104.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied Linear Regression Models (3rd ed.). Chicago, IL: Irwin.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). Fort Worth, TX: Harcourt Brace College.

Peng, C. Y. J., Manz, B. D., & Keck, J. (2001). Modeling categorical

predictor variables by logistic regression. American Journal of Health Behavior, 25, 278-284.

Press, J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, 73, 699-705.

Rice, J, C. (1994). Logistic regression: An introduction. In B. Thompson (Ed.), Advances in social science methodology (Vol.3, pp. 3-28). Greenwich, CT: JAI Press.

Schiel, J. L., & King, J. E. (1999). Accuracy of course placement validity statistics under various soft truncation conditions (ACT Research Report No. 99-2). Iowa City, IA: ACT.

SPSS, Inc. (1999). SPSS Regression Models 9.0. Chicago: Author.

Stevens, J. (1996). Applied multivariate statistics for the social sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Thompson, B. (1989). Editorial: Why won't stepwise methods die? Measurement and Evaluation in Counseling and Development, 21, 146-148.

Thompson, B. (1990). MULTINOR: A FORTRAN program that assists in evaluating multivariate normality. Educational and Psychological Measurement, 50, 845-848.

Thompson, B. (1999). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. Exceptional Children, 65, 329-337.

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

Appendix A

SAS Syntax for Obtaining Logistic Regression Results (for versions prior to 6.06):

```
PROC CATMOD;
 DIRECT mpg engine horse weight accel;
 MODEL origin=mpg engine horse weight accel / ML NOGLS;
```

SAS Syntax for Obtaining Logistic Regression Results (for version 6.06 or later):

```
PROC LOGISTIC;
 MODEL origin = mpg engine horse weight accel;
```

SPSS Syntax for Obtaining Logistic Regression Results

```
LOGISTIC REGRESSION VAR=origin
  /METHOD=ENTER mpg engine horse weight accel
  /CLASSPLOT
  /PRINT=GOODFIT CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Appendix B

SPSS Syntax for Calculating R^2

```
COMMENT Calculate logistic regression estimates saving predicted probabilities
.
LOGISTIC REGRESSION VAR=origin
  /METHOD=ENTER mpg engine horse weight accel
  /SAVE PRED
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

COMMENT Correlate criterion variable with predicted probabilities .
CORRELATIONS
  /VARIABLES=origin pre_1
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Appendix C

SAS Syntax for Generating a Best Subsets Logistic Regression Which Includes Mallow's C_p

```
* Run logistic regression for full model, saving predicted
  probabilities (pred);

  PROC LOGISTIC;
   MODEL origin = mpg engine horse weight accel
   OUTPUT out = output1
            p = pred;

* Define two new variables: z and u;

  z = log(pred / (1 - pred)) + ((sex - pred) / (pred * (1 -
      pred)));
  u = pred * (1- pred);

* Run linear regression to obtain Cp;

  PROC REG;
   MODEL z = mpg engine horse weight accel
   / SELECTION = RSQUARE CP;
   WEIGHT u;
```

Appendix D

Abbreviated Output from SAS Linear Regression "Best Subsets" Procedure

Number in Model	R-Square	C(p)	Variables in Model
1	0.0898	45.5885	ENGINE
1	0.0197	78.9335	WEIGHT
1	0.0149	81.2097	MPG
1	0.0075	84.7290	HORSE
1	0.0011	87.7647	ACCEL
2	0.1475	20.2030	ENGINE WEIGHT
2	0.1354	25.9147	ENGINE HORSE
2	0.0919	46.6025	MPG ENGINE
2	0.0911	46.9852	ENGINE ACCEL
2	0.0229	79.4234	MPG WEIGHT
2	0.0209	80.3667	WEIGHT ACCEL
2	0.0198	80.8667	HORSE WEIGHT
2	0.0153	83.0233	MPG ACCEL
2	0.0152	83.0613	MPG HORSE
2	0.0077	86.6224	HORSE ACCEL
3	0.1861	3.8503	ENGINE HORSE WEIGHT
3	0.1704	11.2784	ENGINE WEIGHT ACCEL
3	0.1475	22.1665	MPG ENGINE WEIGHT
3	0.1460	22.9104	ENGINE HORSE ACCEL
3	0.1363	27.5221	MPG ENGINE HORSE
3	0.0932	47.9802	MPG ENGINE ACCEL
3	0.0241	80.8221	MPG HORSE WEIGHT
3	0.0238	80.9578	HORSE WEIGHT ACCEL
3	0.0236	81.0750	MPG WEIGHT ACCEL
3	0.0153	84.9961	MPG HORSE ACCEL
4	0.1897	4.1272	MPG ENGINE HORSE WEIGHT
4	0.1870	5.4353	ENGINE HORSE WEIGHT ACCEL
4	0.1704	13.2783	MPG ENGINE WEIGHT ACCEL
4	0.1493	23.3292	MPG ENGINE HORSE ACCEL
4	0.0305	79.8169	MPG HORSE WEIGHT ACCEL
5	0.1900	6.0000	MPG ENGINE HORSE WEIGHT ACCEL

Table 1

Excerpt from *Cars* Data Set

Case #	origin	mpg	engine	horse	weight	accel
1	1	18	307	130	3504	12
2	1	15	350	165	3693	12
3	1	18	318	150	3436	11
4	1	16	304	150	3433	12
5	1	17	302	140	3449	11
6	1	15	429	198	4341	10
7	1	14	454	220	4354	9
8	1	14	440	215	4312	9
9	1	14	455	225	4425	10
10	1	15	390	190	3850	9
11	1	15	383	170	3563	10
12	1	14	340	160	3609	8
13	1	15	400	150	3761	10
14	1	14	455	225	3086	10
15	0	24	113	95	2372	15
...	...	...	...	...	...	...
391	1	31	119	82	2720	19

Note. origin = country of origin (coded 0 = European/Japanese, 1 = American); mpg = miles per gallon; engine = engine displacement (cu in); horse = horsepower; weight = vehicle weight (lbs); accel = time to accelerate from 0 to 60 s.

Table 2

Summary of Linear Regression Results for Predicting Country of Origin (N = 391)

Variable	b	SE b	β	t	prob	r_s
(constant)	.992	.279		3.554	.000[a]	--
mpg	-.013	.004	-.217	-3.252	.001	-.802
engine	.005	.001	1.087	9.098	.000[a]	.931
horse	-.008	.001	-.600	-5.441	.000[a]	.694
weight	.000	.000	-.089	-.734	.463	.853
accel	-.005	.010	-.031	-.518	.605	-.371

Note: r_s = structure coefficient obtained by correlating the predicted probabilities ($\hat{Y}$ scores) with predictor variable scores. Mult R^2 = .495.

[a]In some statistical packages, prob < .0005 is represented by .000.

Table 3

Summary of Discriminant Analysis Results for Predicting Country of Origin (N = 391)

Variable	Function[a]	Stdzd Function[b]	r_s
(constant)	1.516	--	--
mpg	-.056	-.357	-.690
engine	.021	1.642	.875
horse	-.031	-1.048	.565
weight	.000	-.143	.758
accel	-.022	-.059	-.273

Note: r_s = structure coefficient. Canonical R^2 = .495 (Wilks' λ = .505).

[a]Function coefficients are analogous to regression b weights.

[b]Standardized function coefficients are analogous to regression β weights.

Table 4

Summary of Logistic Regression Results for Predicting Country of Origin (N = 391)

Variable	b	SE b	Wald's t	prob	OR[a]
(constant)	-.114	3.082	.001	.971	--
mpg	.035	.039	.808	.369	1.035
engine	-.105	.106	42.661	.000[b]	.900
horse	.050	.022	5.222	.022	1.052
weight	.004	.001	10.876	.001	1.004
accel	-.028	.105	.072	.789	.973

Note: Null model -2LL = 517.724, full model -2LL = 212.659, difference $\chi^2(5)$ = 305.065, $p < .001$.

[a]OR = odds ratio; also denoted exp(b). Confidence intervals for the OR estimates can be obtained in both SAS 8.0 and SPSS 10.1 (not listed here).

[b]In some statistical packages, prob < .0005 is represented by .000.

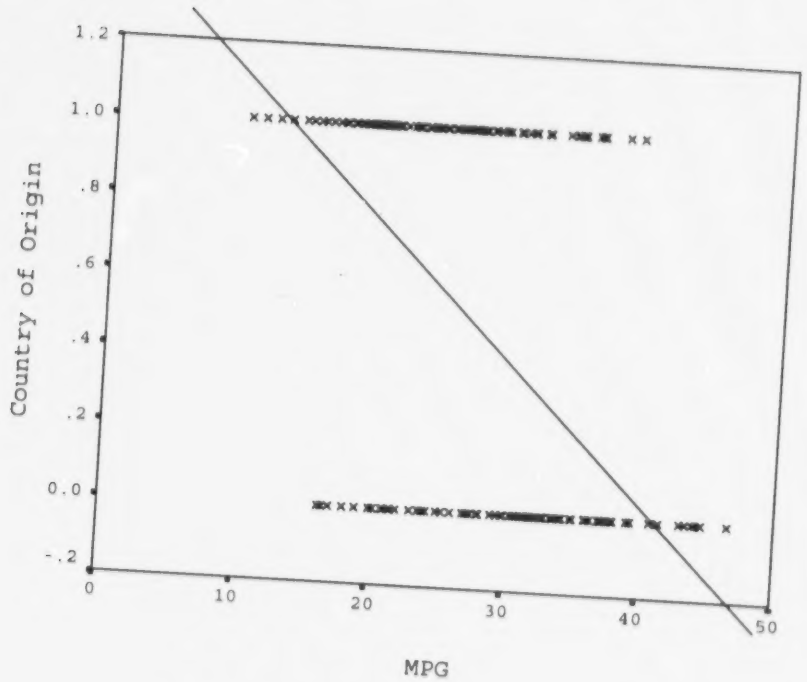
Figure Captions

Figure 1. Scatterplot of mpg with origin depicting a nonlinear relationship.

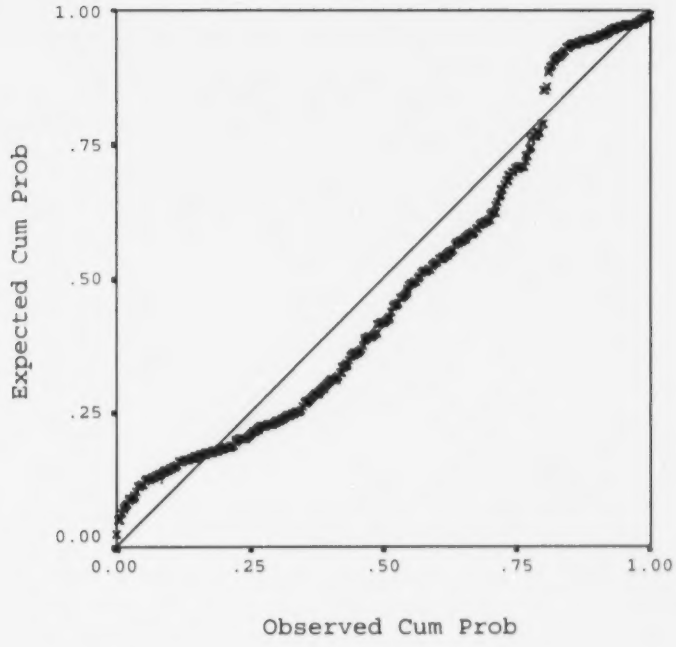
Figure 2. P-P plot of observed versus cumulative probabilities illustrating the non-normality of errors.

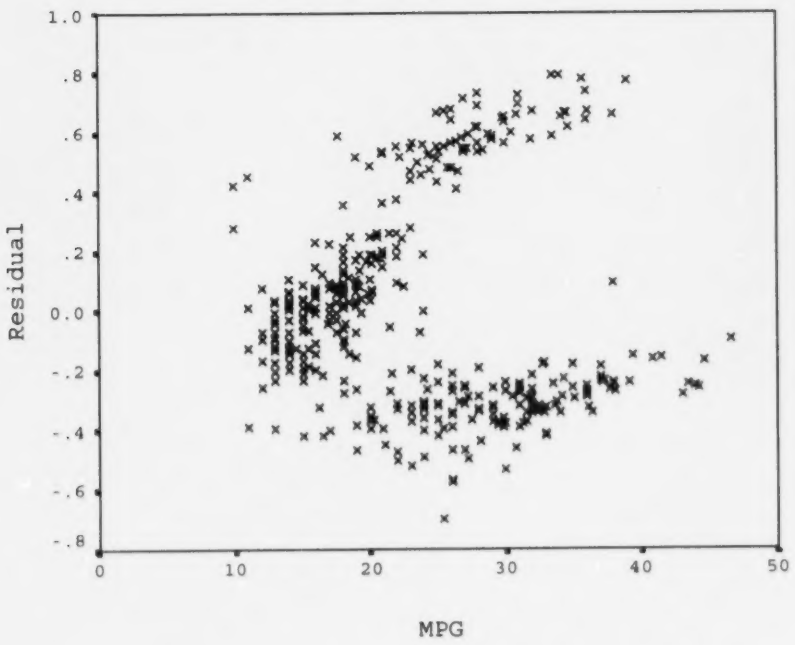
Figure 3. Scatterplot of mpg with residuals illustrating heteroscedasticity.

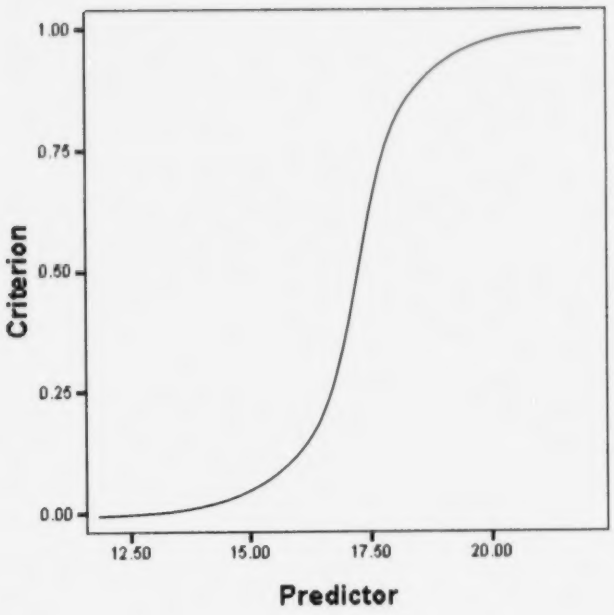
Figure 4. A typical single-predictor logistic regression curve (ogive function).



BEST COPY AVAILABLE


1.00
.75
.50
.25
0.00
Expected Cum Prob
0.00
.25
.50
.75
1.00
Observed Cum Prob


1.0
.8
.6
.4
.2
0.0
-.2
-.4
-.6
-.8
Residual
0
10
20
30
40
50
MPG


1.00
0.75
0.50
0.25
0.00
Criterion
12.50
15.00
17.50
20.00
Predictor

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM033785

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Logistic Regression: Going Beyond Point-and-Click	
Author(s): Jason E. King	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ↑ ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑ ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑ ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Jason King	Printed Name/Position/Title: Jason King	
Organization/Address: Baylor College of Medicine 1709 Dryden, Suite 534, Houston, TX 77030	Telephone: 713-798-8547	FAX: 713-798-6516
	E-Mail Address: jasonk@bcm.tmc.edu	Date: 4/4/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)